

Measuring the Efficiency of Prescribing by General Practitioners

J. M. BATES, D. BAINES and D. K. WHYNES
University of Nottingham

Data envelopment analysis (a mathematical programming technique) has often been applied to measuring the efficiency with which outputs are produced. The technique derives efficient combinations of outputs for given inputs: constant returns to size may be assumed or one may choose to examine whether decreasing or increasing returns hold true. An analysis of the cost of prescribing drugs for 106 general practices in the Lincolnshire Health Authority for the year 1993/1994 reveals the statistical problems that are encountered in applying this technique.

Key words: data envelopment analysis, Health Service

OBJECTIVES OF STUDY

The main objective of this study is to illustrate the nature of the problems arising when attempting to assess the efficiency with which the prescribing of drugs is achieved by general practitioners in the UK National Health Service. The National Health Service is concerned with the efficiency with which medical services are provided, and an analysis of the costs of prescribing provides an introduction to what can be done, albeit in just one small part of the overall system. Though there has been an attempt to create some markets in the provision of health outputs in the UK in recent years, the absence of monetary outputs makes such a task less than straightforward. After examining the theoretical ways of analysing efficiency we will consider ways in which output can be measured. Then an empirical assessment will be made of the efficiency of prescribing of 106 general practices in Lincolnshire, for which data of the characteristics of the patients were available, as were prescription costs.

There are two basic approaches used by analysts in measuring efficiency. One specifies a 'production function' relating outputs to inputs in a particular mathematical form. Having specified the mathematical form a relationship is then fitted to observed data of outputs and inputs. This technique is referred to as a 'parametric' method, though strictly it is a method which utilises a pre-determined statistical distribution. At one time methods such as ordinary least squares were employed in which the data were fitted on average by the function specified. More recently, econometricians have been concerned to specify not just a term for the basic variability in the outcomes, but also a one-sided inefficiency term, these error terms being specified by statistical distributions, and estimated by maximum likelihood methods. See Bauer¹ for a review article. The alternative approach is to use data envelopment analysis (DEA), a particular application of mathematical programming. The DEA technique has been developed to analyse questions of general substitutability between outputs and inputs: its use of the ratio of (weighted) outputs to (weighted) inputs enables any number of outputs to be considered. It is a distribution-free method (commonly called 'nonparametric') in which the efficiency frontier is determined by the data. As such, the data are allowed to 'speak for themselves' and the imposition of a specific production function is avoided. This has much appeal to many practitioners, but the analyst is offered many choices with regard to assumptions about returns to size, and frequently about the data which are to be utilised. Indeed, the contention of this paper is that in many studies involving health outputs, the number of possibly relevant factors describing the social and economic background of patients is very large. The objective of this paper is to examine the problems encountered when such circumstances arise. In particular, though a case will be made for utilising DEA in many of the situations, it will be argued that the choice of assumption about returns to size is important, for it can have a large influence on the nature of the results obtained.

Correspondence: J. M. Bates, Department of Economics, University of Nottingham, Nottingham NG7 2RD, UK

GENERAL ARGUMENTS SUGGESTING THE USE OF DATA ENVELOPMENT ANALYSIS

Analysis of the efficiency with which units deliver outputs has for many years been considered as best done by means of Data Envelopment Analysis. Farrell² identified two particularly important reasons for this: first, it is necessary to exclude points which are not on the production frontier, and second the ability to consider many outputs. These ideas, developed by Charnes, Cooper and Rhodes³ into what is now known as Data Envelopment Analysis, can be illustrated easily by means of an example. Consider an example in which inputs are combined in strict proportions. Figure 1 depicts the output levels of two commodities for a given level of inputs. An analysis which gave weight to all the points would fail to find the efficient boundary. The method works most reliably when there are a large number of producers and it points up non-efficient producers (who are revealed as not on the efficiency boundary). The DEA method is not restricted to constant levels of inputs and forms the ratio of (weighted) outputs to (weighted) inputs in order to consider many outputs and inputs.

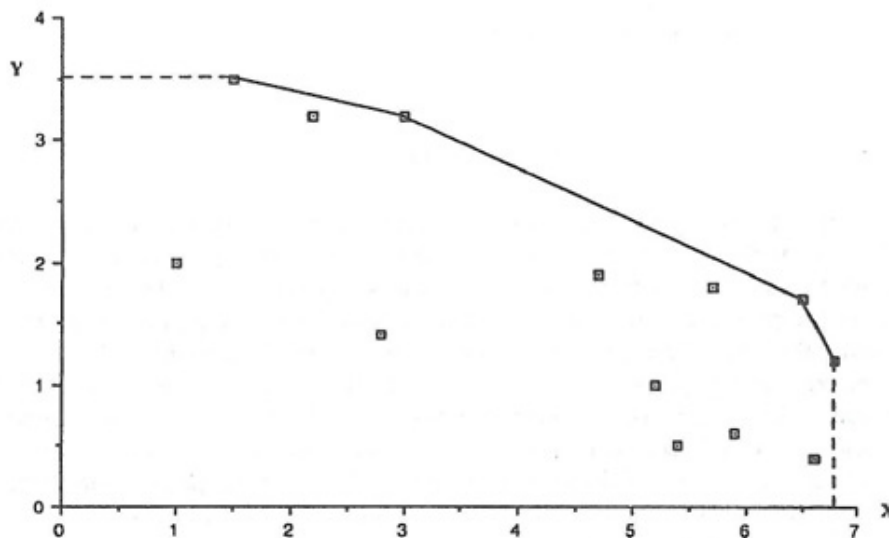


FIG. 1. Outputs of two commodities for a given level of inputs.

These arguments were so important as to lead most practitioners in this area to take the DEA approach as the obvious starting point for analysis of measures of efficiency. With the development of econometric methods of frontier analysis, the arguments are less conclusive. Econometric methods now utilise two 'error' distributions, one relating to general variability of outcomes, and the other being a one-sided inefficiency term: see Jondrow, Lovell, Materov and Schmidt⁴. But problems of estimating relationships when multiple outputs are endogenous (and hence correlated with the error terms) have yet to be solved using these econometric methods.

CHOOSING THE APPROPRIATE DEA MODEL

The prescription which follows is based on the algebra given in Banker, Charnes and Cooper⁵, and in Kooreman⁶. Denote outputs of decision making unit (DMU) k by q_{ik} and inputs by m_{jk} , where in our example the letter i specifies the particular outputs and the letter j the inputs. In deriving a measure of efficiency for DMU a , say, the use made of Data Envelopment Analysis in this paper is to determine the weights w_{ia} and v_{ia} from the program:

$$\text{Maximise } \left\{ \sum w_{ia} \times q_{ia} / \sum v_{ia} \times m_{ja} \right\} \quad (1)$$

$$\text{subject to } \left\{ \sum w_{ia} \times q_{ik} / \sum v_{ia} \times m_{jk} \right\} \leq 1, \quad \text{for all } k. \quad (2)$$

The usual conditions $w_{ik}, v_{ik} \geq 0$ apply.

These hyperbolic functions can be simplified. The trick is to find multipliers which make the denominators unity for each DMU.

Multiply the numerator and denominator of equations (1) and (2) by t_k , which are yet to be determined.

Now set the denominator of the expressions in equation (2) to be equal to 1, finding the values of t_k which yield this condition. Now it remains to find which of the k DMUs has the highest efficiency; set this efficiency to unity, and solve the level of efficiency for DMU a .

For the constant returns to scale case considered so far, since each DMU is restricted to have an efficiency < 1 , we impose the restriction

$$\sum w_{ia} \times q_{ik} < \sum v_{ia} \times m_{jk} \quad (3)$$

Where other scale assumptions are mooted

$$\sum w_{ia} \times q_{ik} + s < \sum v_{ia} \times m_{jk} \quad (4)$$

where, for example, s being limited to nonpositive values allows constant or decreasing returns to scale. In a similar way, s being semipositive allows constant or increasing returns to scale. For the algebra relating to different assumptions see the discussion under multiplier problems for model DI_p on pages 13–15 of the survey article by Seiford and Thrall⁷, and pages 16–27 for an illustration of the assumptions.

DESCRIPTIVE VARIABLES

The number of possibly relevant descriptive factors concerning the social and demographic background of patients is very large. Two features may require care. They are:

- (a) the number of factors to consider;
- (b) the treatment of returns to scale.

The number of factors that could be considered is indeed large, and many could be relevant in a study of health outcomes. Some of them may have such a large impact on general health, and hence expenditures, that it would be inadmissible to ignore them. One problem which frequently arises is how many of these factors to include in the study. Where a characteristic has no effect on costs or outputs its inclusion would not be appropriate; for, if it were included, there would be much scope for increases in the proportion of DMUs which would be classified as efficient. Many studies have been undertaken in which the first stage is to assess which of the characteristics have an effect on outcomes, so that those which appear of little importance are omitted before analysis is undertaken. For example, Kooreman⁶ undertook a DEA analysis of nursing home care using 4 output measures and 6 inputs: he subsequently examined if the efficiency measures were related to 24 descriptive variables. His study gave no hint of any problems in the selection of the variables. In contrast, in a study by Salinas-Jiménez and Smith⁸ which used 3 inputs in a study of primary health care, they acknowledged that the 'wide extent of ... possible additions to the list of inputs (suggests) caution in interpreting the results'. For our study it seemed advisable to specify a formal procedure for determining if a variable should be included.

First, the general procedure. Where many variables are of interest, the procedure to adopt has much similarity with the stepwise regression approach. First, choose the variable which is most correlated with the (latest) efficiency measure, provided its contribution is significantly different from zero. In examining the degree of significance, it is important to recall that a search for significance of k variables amongst c candidates should have the significance level modified. Lovell⁹ notes that, where the explanatory variables are orthogonal, the true value for α , the significance level, is related to the supposed level $\hat{\alpha}$ by the equation:

$$\alpha = 1 - (1 - \hat{\alpha})^{c/k} \quad (5)$$

The variables will not be entirely orthogonal, so this equation might make the degree of significance appear less than it is, but its use mitigates the dangers of data mining.

Next, perform another DEA exercise, generating revised efficiency measures. Then choose as the next variable to introduce the one with the highest correlation with the revised efficiency measure, provided once again that this correlation is significantly different from zero.

One feature requires some comment. It concerns the possible exclusion of a variable which had earlier been selected. If, at any stage, the correlation between the measure of efficiency (when excluding the variable under investigation) and a variable is not 'significant', then one can eliminate that variable from the list of explanatory variables. This corresponds to the usual treatment in stepwise regression.

There is also the choice to be made about the kind of 'scale' treatment to postulate. An assumption that there are economies of scale would imply that a DMU with the lowest level of input for one particular variable would inevitably obtain an efficiency score of unity, irrespective of its outputs. In the example chosen by Seiford and Thrall⁷, only one of the five DMUs is efficient when the constant returns to scale assumption is utilised (see solid line), but this increases to 3 in the constant or decreasing returns to scale assumption (dotted line). Figure 2 depicts the five DMUs, with 'efficient' DMUs being on the boundary.

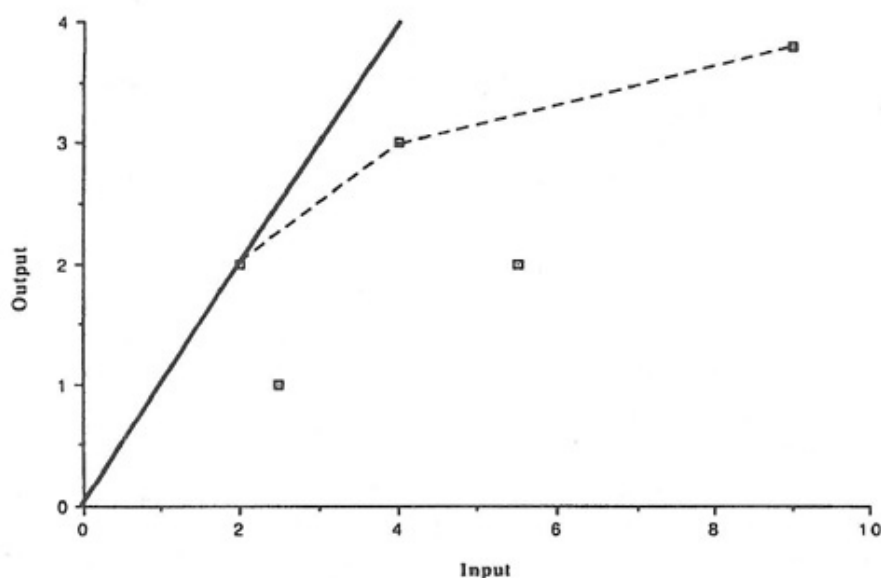


FIG. 2. Illustration of the constant or decreasing returns assumption.

Note that the choice of model may have considerable impact on the number of DMUs regarded as efficient: this may be of considerable importance when considering 'characteristics' of the DMUs, for there are many social and economic characteristics which might be thought to be relevant in determining the costs of achieving particular output levels in health studies. Where the concern is to identify efficient or inefficient DMUs, it seems better to assume constant returns to scale. As will be seen later, constant returns to scale is a plausible assumption in our example. It may not be the fault of the 'manager' of an individual DMU, but a failure to obtain the appropriate scale of unit is a form of inefficiency.

FITTING THE DEA MODEL

As a measure of input the levels of prescribing expenditure was used.

At the outset it was crucial to decide the choice of output measure. Many would argue that health gain, usually measured in Quality Adjusted Life Years (QALYs) is an appropriate measure of output: but, since QALYs cannot be measured for our data, an alternative measure of output is required.

Another possibility would be to use standardised mortality rates. With small numbers of deaths arising in many of the practices in our sample there was a danger that the outcomes could be subject to a large proportionate variation in the year analysed, and hence would be less than fully reliable. Also, the difficulties of precluding the effects of other factors has been noted by Marinker¹⁰. So we did not seek the calculation of standardised mortality rates.

In the end we chose as an initial measure of output the number of patients treated. Whynes, Baines and Tolley¹¹ tested whether certain practice and patient characteristics influenced costs per 'ASTRO-

PU' (age, sex, and temporary resident originated prescribing unit) patient. Age, sex and temporary residents were therefore factors: other significant factors were the number of night visits made per patient, the proportion of patients with payment exemptions, whether the practice had fund-holding status, and the proportion of medicines prescribed generically. Unemployment and the number of patients per general practitioner had no significant impact on prescribing costs.

Fund-holding status and the proportion of medicines prescribed generically are factors of a 'management' kind. Apart from the levels of expenditure 'management' indicators can safely be ignored, if all that is required is a measure of performance; for, it is a choice of the practice as to how it organises its affairs in terms of number of partners, whether to be a fund-holder, etcetera.

To perceive what is happening the analysis commences by consideration of the fundamental ratio of patients to cost. The relationship between these two variables can be seen in Figure 3. The percentage efficiency scores are given in column 1 of Table 1.

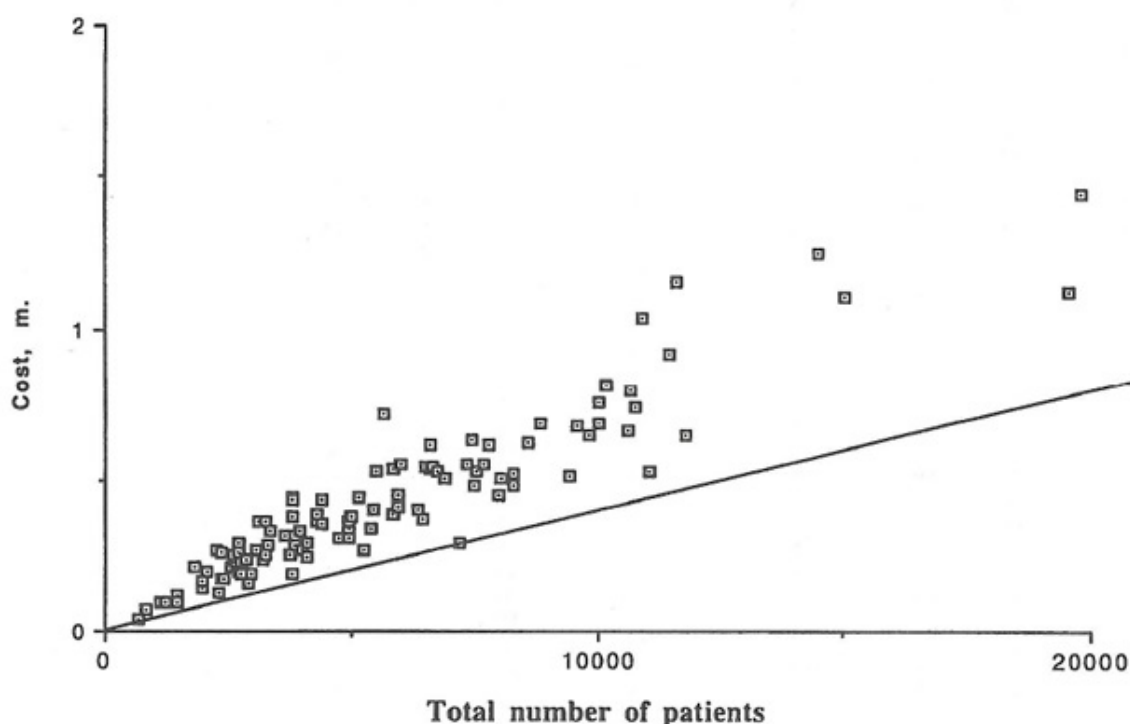


FIG. 3. Cost and number of patients.

Whynes¹² suggests that larger practices may be more efficient by being able to negotiate from strength with health providers, or through being able to spread certain fixed administrative costs, but there is no suggestion that there are economies in the provision of drugs. The graph certainly gives no hint of any economies with regard to size, the degree of inefficiency being unchanged as the number of patients increases: constant returns to scale have therefore been assumed in the analyses undertaken in this work. Furthermore, if there were economies of scale, then any failure to be at the correct size would be a kind of inefficiency. Points on Figure 3 above the solid line represent inefficiencies, when utilising costs as the sole input. The efficiency measure obtained at this first stage was given by

$$\text{Efficiency} = n/e \quad (6)$$

where n is the actual number of patients in the given practice, and e is the expected number of patients for any given total cost, as measured by the line fitted.

Subsequently the output measure was modified to distinguish patients with different characteristics such as age, (the older patients having more health needs, on average) those exempted from payment for prescriptions on grounds of medical need, and the number of temporary residents treated by the practice. These characteristics cannot be ignored; they appear to influence the cost of treating patients and to ignore their effect would unfairly bias the results against those practices having relatively poor scores on these measures.

TABLE 1. *Percentage efficiencies—1 input, and up to 4 outputs*

Outputs used					Outputs used				
Practice number	[1] List size	[2] Old young	[3] Old, exempt young-exempt	[4] Old, exempt young-exempt, temp res	Practice number	[1] List size	[2] Old young	[3] Old, exempt young-exempt	[4] Old, exempt young-exempt temp res
1	100	100	100	100	54	52	67	69	69
2	42	59	63	100	55	64	64	69	69
3	34	47	50	100	56	53	62	69	69
4	79	94	99	99	57	49	66	68	68
5	75	90	99	99	58	54	62	67	67
6	82	85	97	97	59	49	56	67	67
7	76	100	96	96	60	56	60	66	66
8	43	43	96	96	61	47	48	66	66
9	58	84	95	95	62	64	64	65	65
10	54	55	93	93	63	53	62	65	65
11	84	90	85	85	64	45	56	65	65
12	51	90	84	84	65	60	63	64	64
13	61	62	84	84	66	34	63	64	64
14	42	65	68	84	67	55	61	64	64
15	38	39	83	83	68	47	59	64	64
16	64	78	81	81	69	49	59	64	64
17	74	86	80	80	70	50	55	64	64
18	73	73	79	80	71	62	64	63	63
19	37	75	68	80	72	49	60	63	63
20	71	75	79	79	73	49	60	63	63
21	74	77	78	78	74	45	51	60	63
22	67	76	78	78	75	47	56	62	62
23	55	67	78	78	76	55	64	61	61
24	41	51	78	78	77	61	61	61	61
25	61	71	77	77	78	52	59	61	61
26	62	69	76	76	79	48	57	61	61
27	51	66	76	76	80	46	56	61	61
28	51	60	61	76	81	54	56	61	61
29	64	67	76	75	82	54	54	61	61
30	69	70	75	75	83	46	56	59	59
31	47	48	75	75	84	43	52	59	59
32	56	71	74	74	85	47	61	58	58
33	64	66	74	74	86	58	59	58	58
34	61	66	74	74	87	57	58	58	58
35	63	70	73	73	88	50	57	58	58
36	54	68	73	73	89	53	56	58	58
37	55	62	73	73	90	43	53	58	58
38	49	49	73	73	91	41	52	58	58
39	69	73	72	72	92	53	57	57	57
40	70	72	72	72	93	44	46	57	57
41	59	67	72	72	94	41	50	56	56
42	58	67	72	72	95	55	55	55	55
43	55	66	72	72	96	48	64	54	54
44	46	47	72	72	97	41	49	54	54
45	32	51	52	72	98	42	49	54	54
46	64	70	71	71	99	47	48	54	54
47	64	68	71	71	100	36	46	54	54
48	50	68	71	71	101	52	52	52	52
49	59	61	71	71	102	35	41	52	52
50	56	57	71	71	103	48	48	49	49
51	57	67	70	70	104	44	46	47	47
52	56	65	70	70	105	34	40	46	46
53	48	59	68	70	106	35	35	42	42

Additional factors will now be considered: the analysis proceeds by adding one variable at a time. Age was the next variable introduced: this variable was chosen since it yielded a 'better' explanation than any other single variable of the differences in 'efficiency'. Information was available on just 3 age groups, those under 65, those aged between 65 and 74, and those 75 and above. When an analysis was undertaken on how age influenced cost the relationship obtained was:—

$$\text{Total Cost} = 22.02 (\text{age} \leq 65) + 0.55 (\text{age } 65-74) + 518.5 (\text{age} \geq 75) \quad (7)$$

(5.59) (114.5) (198.5) $R^2 = 0.93$

The figures in brackets represent standard errors. The choice of the proportion aged 75 and over as the variable to include is based on the insignificant coefficient for the 65–74 age group, the coefficient being not significantly different from the less than 65 group coefficient.

The 'natural' way in which to incorporate this variable into the DEA model is to distinguish the number of patients aged 75 and over (the 'elderly' patients) and have two outputs, the number of elderly patients and the other patients (here referred to as 'young'). One would thus be examining the costs of the 106 practices against the number of young patients treated and against the number of elderly patients. The results of analysing the problem in this way are presented in Figure 4, with the efficiency scores given in column 2 of Table 1. Figure 4 depicts the efficiency frontier by the solid straight line. The dotted lines joining A and B to the axes and those traditionally used. Thus point C is recorded as being 90.3% efficient, even though it is true that, if there were more 'young' in that practice, DMU B has shown that they could be treated at no extra cost. Point D is unambiguously 90% efficient, being 90% of the distance from the origin to the efficiency frontier.

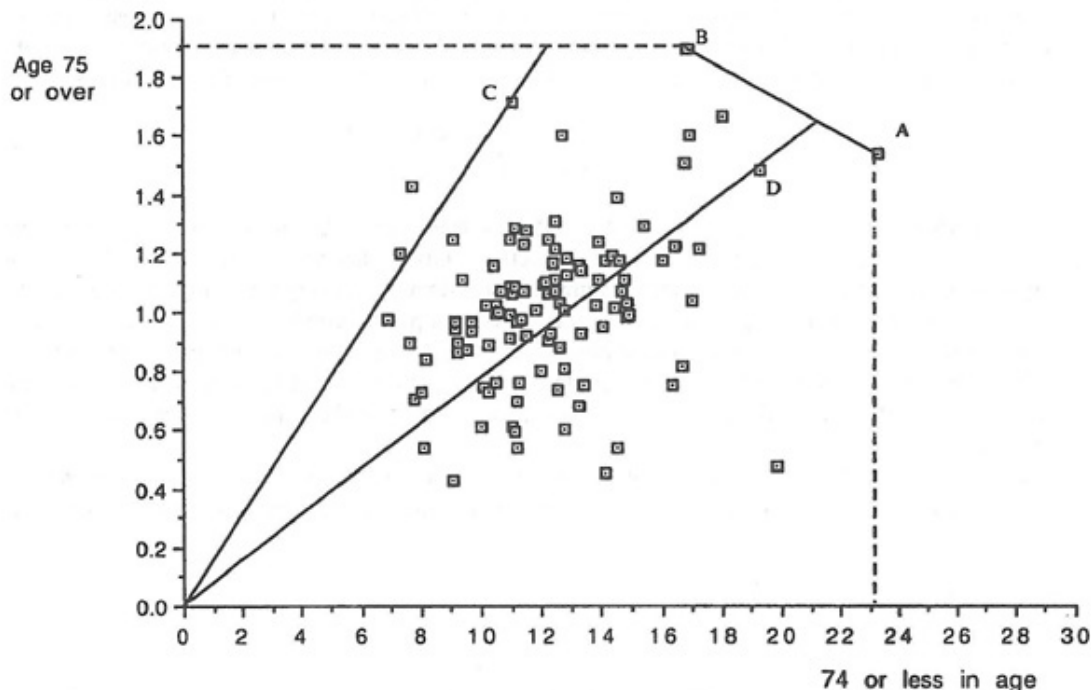


FIG. 4. Number of patients per 1000.

When following this approach there is often a possibility of obtaining many 'efficient' practices. As it happened, there are two efficient points only on the graph (denoted as A and B), so this feature does not arise in this example. So, where the implied substitution of old for young in cost terms might have varied throughout the graph, this was not so here. The substitution ratio of each elderly patient costing 17.76 times that of a 'young' patient may appear large but is not inconsistent with a linear regression of the costs which gave the equation

$$\text{Cost} = 36.03 \text{ (ages 0–74)} + 526.79 \text{ (age 75 and above)} \quad (8)$$

(5.57) (66.63) $R^2 = 0.91$

Here the ratio of costs is 14.6:1 for the 'elderly' in comparison with the 'young'.

FURTHER ANALYSIS

Measures of efficiency were derived using the over 75s and the under 75s as the two measures of output (see Figure 4). The set of measures of efficiency was then compared with the variables not yet included in the analysis, namely the proportion of 'exemptions' in each practice's list and the ratio of

temporary residents in the list size. The term 'exemptions' applies to patients who are exempt from the payment of prescription charges on grounds of permanent ill-health, or who have purchased a 'season ticket' for a period of at least 4 months: patients who are of retiring age do not pay prescription charges, but these are not classified as 'exemptions'.

An analysis of these results showed that the lower the measure of efficiency the greater the proportion of patients who were exempt from payment of prescription charges: the relationship had coefficients significantly different from zero. This result is not surprising: patients classified as 'exempted' from paying for prescriptions are principally those who have a long-term need for drugs arising from some illness but who are not exempt on the grounds of age. In contrast, there was less correlation between the measures of efficiency and the number of temporary residents.

At this point the traditional DEA analysis of including another output variable was undertaken. It was necessary to deduct the number of exemptions from the number of patients below 65 years of age in order to have three distinct output groups. The results of this analysis are given in column 3 of Table 1.

Other variables were now considered. The only one which was associated with the latest measure of efficiency was the number of temporary residents. The latest efficiency measures was regressed on the proportion of temporary residents per each practice's adjusted list size, the adjustment being to weight the elderly and exempt patients more highly than the 'young' patients. The resulting equation was

$$\text{Efficiency} = 69.47 - 41.14 (\text{temp res/L}) \quad (9)$$

(12.05) (21.74)

The coefficient was not (quite) significantly different from zero. One feature was apparent. The regression result is based heavily on just a few practices. Given that some practices are located in holiday resorts the imbalance in the number of temporary residents is not surprising: Figure 5 depicts the relationship between cost and the ratio of temporary residents per adjusted list size. It is evident how little variability there is in the majority of practices. Though the effect of temporary residents on prescribing charges cannot be reliably estimated, logic states that costs must be increased. The variable is therefore correctly included as an explanatory factor. So a further stage in the analysis was undertaken.

Relating cost to the four output measures (the number of elderly, exempt, young less those exempt, and temporary residents) gave a final analysis, the results of which are reported in the final column of

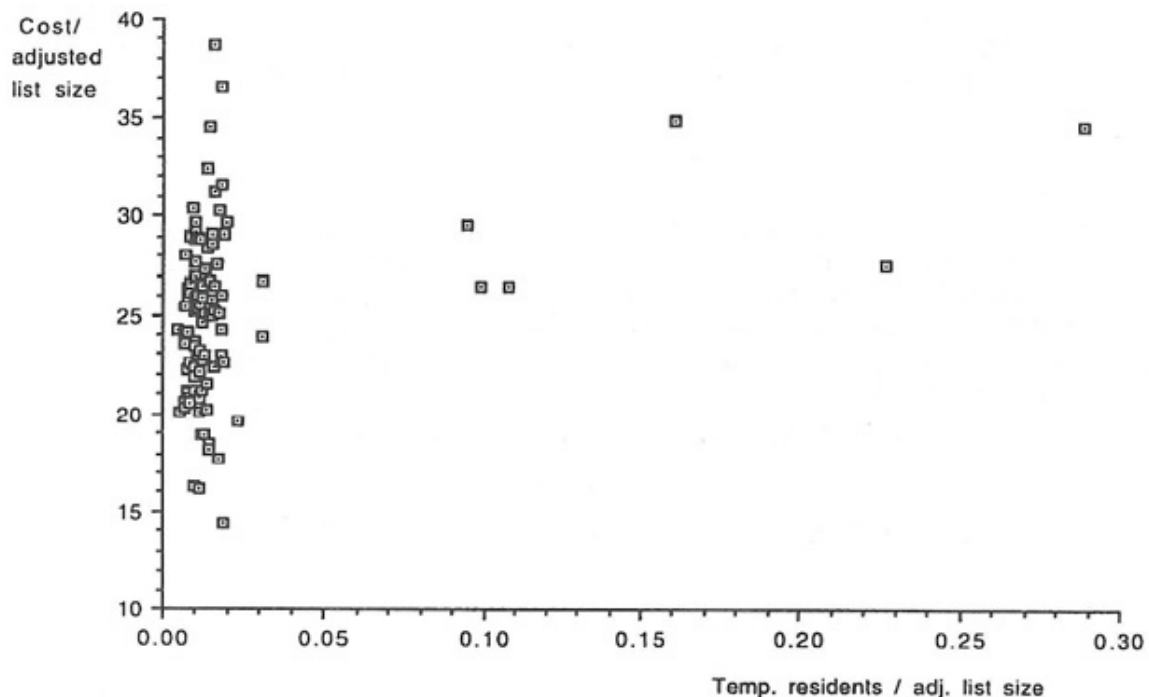


FIG. 5. Temporary residents and cost.

Table 1. Yet again there are few (this time 3) practices classified as 100% efficient. It might be instructive to ask what features these 3 practices share. From the statistical information available we know that just one of the three was a fund-holder; the other two operated a practice formularily, ie only drugs contained in a list specified by the practice being prescribable. Just under half the practices operated a prescribing formularily: eight of the ten practices ranked as most efficient operated a practice formularily, but then so did five of the ten lowest ranked practices. Also, not one of the 19 fundholders was in the lowest quarter of Table 1, which is ranked by the numbers in the final column, yet it was possible for non-fundholders to be efficient. These results suggest that there is more than one route to efficiency in prescribing.

Though the measure of efficiency is much altered for a few practices, a comparison of the four sets of results recorded in Table 1 indicates that, for most of the practices, the difference in the measure of efficiency varies little as the additional factors are considered. For most of the practices in the lower half of the table, there is little doubt that their efficiency is 70% or less.

CONCLUSIONS

The study has indicated some features which commonly arise in attempting to assess the efficiency with which health outputs are achieved. The first concerned measures of output, where we chose to examine the number of patients of different kinds treated. It would be instructive to see if other possible measures of output gave similar results. The DEA analysis which was undertaken has proved to be a helpful starting point for the investigation, but certain statistical choices were available. The approach adopted here assumed constant returns to size, which enabled some understanding of relationships between outputs and inputs to be derived. But it has shown that, even with as many as 106 practices, it was not possible to derive wholly reliable measures of efficiency when one parameter is based on very few practices, as occurred with the data on temporary residents.

This paper has shown that Data Envelopment Analysis can be used to assess the efficiency of prescribing, though it needs to be undertaken with care.

Acknowledgements—Helpful comments from Gülay Günlük-Şenesen of Istanbul Technical University and anonymous referees are gratefully acknowledged by the authors. The generous provision of data by the Lincolnshire Health Authority is gratefully acknowledged.

REFERENCES

1. P. W. BAUER (1990) Recent developments in the econometric estimation of frontiers. *Journal of Econometrics* 46 (1/2), Annals, 39–56.
2. M. J. FARRELL (1957) The measurement of productive efficiency. *J. Roy. Stat. Soc. A* 120, 253–281.
3. A. CHARNES, W. W. COOPER and E. RHODES (1978) Measuring the efficiency of decision making units. *Eur. J. Opl. Res.* 2, 429–444.
4. J. JONDROW, C. A. KNOX LOVELL, I. S. MATEROV and S. SCHMIDT and S. SCHMIDT (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics* 19, 233–238.
5. R. D. BANKER, A. CHARNES and W. W. COOPER (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Mngt Sci.* 30, 1078–1092.
6. P. KOOREMAN (1994) Nursing home care in the Netherlands: a nonparametric efficiency analysis. *Journal of Health Economics* 13, 301–316.
7. L. M. SEIFORD and R. M. THRALL (1990) Recent developments in DEA: the mathematical programming approach to frontier analysis. *Journal of Econometrics* 46(1/2), Annals, 7–38.
8. J. SALINAS-JIMÉNEZ and P. SMITH (1994) Data envelopment analysis applied to quality in primary health care. University of York, Discussion paper 124.
9. M. C. LOVELL (1983) Data mining. *The Review of Economics and Statistics* 65, 1–12.
10. M. MARINKER, Ed. (1990) *Medical audit and general practice*. British Medical Journal, London.
11. D. K. WHYNES, D. L. BAINES and K. TOLLEY (1996) Explaining variations in general practice prescribing costs per ASTRO-PU (age, sex and temporary resident originated prescribing unit). *British Medical Journal* 312, 488–489.
12. D. K. WHYNES (1993) The NHS internal market: economic aspects of its medium-term development. *International Journal of Health Planning and Management* 8, 107–122.

Received December 1995; accepted April 1996 after one revision